Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Temporal fact verification: Auditing evidence at claim time

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-11-03 · temporal fact verification · temporal fact checking · evidence validity · point in time verification · temporal provenance · historical source verification · bitemporal data · fact verification audit · abstention

Original article: https://reliablecontext.com/articles/temporal-fact-verification-evidence-validity



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes
4. Research approaches and their limits
5. Implementation Considerations and Process Changes
6. A paired temporal evidence fixture
7. Evaluation design
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Conclusion

Executive Summary

Temporal fact verification asks whether evidence supports or contradicts a claim for its specified time, under an explicit evidence-availability policy. A relevant quotation and a correct answer are insufficient if the quotation concerns another interval, a different occurrence, or a version unavailable at the decision cutoff. Existing work establishes several useful components: shared claim-evidence timelines, temporal knowledge-graph reasoning, and event-order reasoning. These are distinct research tasks, rather than interchangeable measurements of temporal reliability. (aclanthology.org) (ris.uni-paderborn.de) (ijcai.org)

The numerical distinctions matter. In **Allein et al. (2023)**, the MultiFC bidirectional long short-term memory (BiLSTM) baseline moves from **55.20% to 64.17% Micro F1** and **32.39% to 47.43% Macro F1**. Those differences are **8.97 and 15.04 percentage points**, explaining the abstract's rounded “9%” and “15%”; they are not relative improvements of those sizes. (aclanthology.org) **Qudus et al. (2023)** report abstract-level gains of **0.13 to 0.15 in area under the receiver operating characteristic curve (AUROC)**, separately from **0.25 to 0.27 in mean reciprocal rank (MRR)**; the displayed AUROC table requires a narrower interpretation discussed below. (ris.uni-paderborn.de) **Barik et al. (2025)** report **85.49 Macro F1** for ChronoFact on ChronoClaims, against **63.42** for TACV. That constructed dataset lacks an insufficient-evidence class, so this result does not establish abstention quality. (ijcai.org)

This report proposes an audit organized around a fixed claim and four evidence clocks: the asserted event or valid interval, source publication and revision, observation or retrieval, and database recording. Claim time is the separate reference against which those clocks are assessed. Bitemporal storage can preserve application-time and system-time histories; it does not itself establish textual entailment or source reliability. (mariadb.com) Revision provenance and archival datetimes provide complementary evidence about versions, rather than substitutes for event time. (www.w3.org) (www.rfc-editor.org)

The recommended decision process keeps **textual support**, **temporal eligibility**, and **verdict** separate. A fictional paired fixture tests stale content, revised sources, recurring events, contradictions, and future publication. Evaluation then reports retrieval coverage, eligibility precision and recall, verdict performance, and abstention at defined coverage levels. Classification with a reject option already treats coverage and accuracy

as separate quantities; this audit extends that distinction to the evidence admitted before classification. (jmlr.org) The result is a reviewable answer about a particular claim, interval, cutoff, and source version, including an explicit unresolved outcome when sufficient admissible evidence is absent.

Introduction and Background

Temporal fact checking is often described as adding dates to verification. The consequential question is more precise: **which proposition was evaluated, for which interval, using which evidence version available under which cutoff?** A system can retrieve a semantically excellent passage and still answer the wrong historical question. Research on shared timelines explicitly incorporates publication dates and textual time expressions, while knowledge-graph research evaluates assertions with temporal qualifications. (aclanthology.org) (ris.uni-paderborn.de)

A useful distinction separates **historical truth verification** from **as-known verification**. The former asks what was true during a target interval, potentially using later historical evidence. The latter asks what the specified evidence could establish by a decision cutoff. These are audit definitions proposed here, not names for a universally adopted benchmark taxonomy. A concrete benchmark-specific policy exists: AVerImaTeC requires evidence published before the claim date. Such a rule must be stated when adopted, rather than silently treated as a requirement of every historical inquiry. (fever.ai)

Evidence sufficiency also differs from answer correctness. The original FEVER task couples a correct verdict with at least one complete annotated evidence set for its joint score. That provides a precedent for testing evidential support rather than awarding credit solely for the label, although it does not supply this report's four-clock audit. (fever.ai) For numerical claims, Strong and Vlachos's TSVer instead uses **304 claims**, **41 fact-checking organizations**, and **400 time series**, illustrating a different evidence object and evaluation setting. (aclanthology.org)

The decision addressed here is whether a time-sensitive claim can be **supported**, **refuted**, or **left unresolved** with temporally admissible evidence. The report compares primary research, defines a reproducible fixture, and separates the measurements needed to inspect failures. Its protocol and invented counts are proposals and worked arithmetic, not results from a deployed verifier or a reproduced cross-paper benchmark.

Key Changes

Replace one timestamp with typed clocks

Replace an undifferentiated “date” with typed assertions. Valid time concerns the world described; transaction or system time concerns the stored record. Oracle distinguishes user-controlled valid time from system-managed transaction time. SQL Server's system timestamps use transaction beginning; event dates need separate evidence. (docs.oracle.com) (learn.microsoft.com)

The valid-time versus transaction-time distinction starts **bitemporal data for fact verification**; a complete evidential test requires more. (docs.oracle.com) For this audit, each evidence item should expose the following fields:

- **Claim interval:** The point or interval quantified by the proposition, including its temporal operator.
- **Asserted event interval:** When the source says the relevant state or event occurred.
- **Publication:** When the identified source version was released, with uncertainty retained. (www.crossref.org)
- **Revision:** When a changed version appeared, linked to the preceding version. (www.w3.org)
- **Observation:** When the evaluator captured or retrieved that particular version.
- **Recording:** When the evaluator's database accepted the observation, distinguished from source time. (learn.microsoft.com)
- **Version identity:** A revision identifier or equivalent reference to the captured content. (www.mediawiki.org)
- **Eligibility status:** Eligible, ineligible, or unknown, with the rule and reason.

The event interval and the source's own lifecycle must remain distinct. The PROV Ontology (PROV-O) provides generation and invalidation times for entities and a revision relation between entities; those properties describe provenance without deciding whether the underlying proposition is true. (www.w3.org)

Preserve interval semantics and uncertainty

The difference between “during,” “throughout,” and “at some point in” changes the required evidence. For “throughout,” an asserted supporting interval must cover the claim interval under the specified boundary convention. Mere overlap is insufficient. OWL-Time, the Time Ontology in Web Ontology Language (OWL), provides relations among instants and intervals, while PostgreSQL documentation makes the difference between inclusive and exclusive range bounds explicit. These distinctions ground the proposed eligibility rules. (www.w3.org) (www.postgresql.org) **Point in time claim verification** specializes the target to a specified instant. **Temporal evidence provenance** preserves the chain linking a conclusion to its source version and typed clocks. (www.w3.org)

An absent endpoint must not silently acquire a substantive meaning. A database range can deliberately omit a bound; that expresses a range extending beyond the other bound. An unknown endpoint in a source may instead mean the evaluator lacks information. The audit should preserve that distinction instead of converting missing metadata into an indefinitely valid fact. (www.postgresql.org)

Likewise, metadata fields require interpretation. DataCite distinguishes resource coverage from collection and describes a Valid date or range separately. Its documentation also says that a published field can fall back to publicationYear when a more specific Issued date is absent. Consequently, apparent day-level precision should not be manufactured from year-level metadata. (datacite-metadata-schema.readthedocs.io) (support.datacite.org)

Treat reproducible storage as necessary infrastructure

MariaDB documents bitemporal tables as combining application-time periods with system versioning; IBM likewise describes a combination of application-period storage and system-period historical tracking. These capabilities address persistence along different temporal axes. They do not establish that an assertion's application-time bounds were correctly extracted or that a captured page was available before a decision. (mariadb.com) (www.ibm.com)

Preserve clocks and their derivations before evaluating eligibility. Reproducibility permits inspection; entailment still needs evaluation. System-versioned history, such as the previous row versions documented by SQL Server, supports examining record changes without replacing the need to identify the underlying source version. (learn.microsoft.com)

Research approaches and their limits

Shared timelines, **temporal knowledge graphs**, and **event timelines** address different tasks. Table 1 separates paper-reported results from the publisher's unevaluated system-of-record perspective.

Table 1 compares evidence objects, temporal representations, and evaluation boundaries.

Approach	Evidence and time representation	Task and dataset	Reported result	Audit limitation
Allein et al., implicit temporal reasoning	Publication dates and extracted text expressions grounded on shared timelines. (aclanthology.org)	Evidence-based classification on MultiFC. (aclanthology.org)	BiLSTM reaches 64.17% Micro F1 and 47.43% Macro F1. (aclanthology.org)	Does not independently label all four evidence clocks.
Qudus et al., TemporalFC	Background knowledge from text and graph representations, including temporal embeddings; graph years differ from retrieval dates. (papers.dice-research.org)	Fact checking and time-point prediction on DBpedia124k and Yago3k. (papers.dice-research.org)	Abstract reports AUROC gains of 0.13 to 0.15 and MRR gains of 0.25 to 0.27. (ris.uni-paderborn.de)	Distinct tasks and corruption procedures; gains cannot be pooled.
Barik et al., ChronoFact	Event timelines; ChronoClaims construction starts from a Wikidata snapshot. (ijcai.org)	Support/refute classification on constructed ChronoClaims. (ijcai.org)	Overall Macro F1 85.49 versus TACV 63.42. (ijcai.org)	No insufficient-evidence class in this dataset. (ijcai.org)
Strong and Vlachos, TSVer	Annotated time frames and time-series evidence. (aclanthology.org)	Numerical claims; 304 claims and 400 series. (aclanthology.org)	Retrieval uses a time-series coverage metric distinct from verdict metrics. (aclanthology.org)	A separate evidence modality, not a replication of the text or graph tasks.
Reliable Context	Publisher's system-of-record position.	A comparable option when the purchasing decision concerns defensible records; not evaluated here as a research verifier.	No comparative performance claim made in this report.	Apply the same externally inspectable evidence audit; do not infer research results from product positioning.

A method's score transfers only with its task definition, construction, baseline, and metric. These comparisons do not establish performance on the complete proposed fixture, including source availability and versions.

Shared timelines do not make publication dates ground truth

Allein and colleagues supply temporal information before or after claim and evidence encoding. Claim dates come from metadata; evidence dates are extracted from snippets or publication expressions. Text expressions are normalized using publication time as document creation time. These are supplied, extracted, or normalized timestamps, rather than uniformly annotated event-validity intervals. (aclanthology.org)

An incorrect publication date can corrupt normalization of expressions such as “last year.” Preserve the raw expression and normalization basis to distinguish extraction errors from classification errors.

Graph validity and event order address different questions

TemporalFC evaluates graph assertions and time-point prediction with multiple background sources. Its experiments retain year information and separately alter assertions or timestamps. Predicting a plausible year for a true assertion is not equivalent to proving that a particular retrieved passage was available at a historical cutoff. (papers.dice-research.org)

ChronoClaims uses a **November 2022 Wikidata snapshot**, generated sentences, and reordered sentences for refutations. Its **40,249 training**, **3,544 validation**, and **3,735 test** instances total **47,528**. Generated event-order contrasts do not establish robustness to page revisions or missing dates. (ijcai.org)

TSVer requires selecting series and time frames. Its coverage measure combines dataset-selection F1 with temporal overlap, separating retrieval from verdict evaluation. The source-version audit proposed here remains an additional test. (aclanthology.org)

Implementation Considerations and Process Changes

Freeze the question before admitting evidence

A temporal fact verification audit begins with a **claim specification** that remains unchanged across paired runs. Preserve original wording, entity, relation, interval, and temporal operator. Wikidata's distinction between start/end qualifiers and point-in-time qualifiers illustrates why a point statement and an interval statement should not be substituted without explanation. (www.wikidata.org)

The audit configuration should declare:

- **Question mode:** Historical truth or as-known verification.
- **Target interval:** Its calendar, time zone, granularity, and boundary convention.
- **Knowledge cutoff:** Required for as-known verification, independently of claim time.
- **Availability test:** Publicly published, actually observed, or recorded inside the evaluated corpus.
- **Source policy:** Accepted source classes and a predeclared method for handling corrections.
- **Conflict policy:** Whether unresolved contradictory evidence forces an unresolved verdict.
- **Completeness requirement:** What combination of evidence suffices for the proposition.
- **Unknown handling:** Which missing metadata prevent eligibility from being established.

An as-known audit of what an organization actually had requires observation or record evidence, not just an earlier publication assertion. Conversely, a study of publicly available evidence can adopt a publication cutoff without pretending that the evaluated organization possessed every published document. AVerImaTeC supplies a concrete pre-claim publication policy, but that benchmark rule does not determine the organization's knowledge boundary. (fever.ai)

Preserve versions before interpreting them

A stable URL alone is insufficient for a revision-sensitive test. MediaWiki's application programming interface (API) exposes revision IDs and timestamps as distinct data, providing an explicit way to request a particular version. PROV-O's revision relation supplies a general provenance connection between revised and original entities. Preserve the evaluated text accordingly. (www.mediawiki.org) (www.w3.org)

Archival evidence requires similar care. Request for Comments (RFC) 7089's Memento-Datetime identifies the represented prior state, and its discussion warns that sparse captures can produce a substantial difference between the requested datetime and the returned state. A requested date must therefore not be reported as the capture's actual date. TimeMaps enumerate versions with archival metadata, while Web ARChive (WARC) records provide capture-context fields such as date, type, and length. (www.rfc-editor.org) (mementoweb.org) (www.loc.gov)

A minimal evidence manifest should retain:

- **Source locator:** The original address and the address of the evaluated version.
- **Exact passage:** The text used for support or refutation, with sufficient context.
- **Clock assertions:** Each timestamp, its source, and whether it was supplied, extracted, or inferred.
- **Capture identity:** The version reference and observation details. WARC capture metadata is one available foundation. (www.loc.gov)
- **Revision relationship:** A pointer to any preceding or correcting version actually available.
- **Eligibility rationale:** The applicable rule, result, and unresolved uncertainty.

Provenance preserves assertions without guaranteeing truth. Eligible sources can be mistaken or conflict.

Separate semantic and temporal decisions

For each candidate, annotate **textual relation** as support, refute, irrelevant, or indeterminate. Separately annotate **temporal eligibility** as eligible, ineligible, or unknown. Only then derive an evidence-conditioned verdict under the declared source and conflict policy. This proposed annotation scheme prevents an attractive answer from silently rehabilitating an inadmissible citation.

The distinction resembles, but does not replicate, evidence-aware benchmark scoring. FEVER requires a complete supporting evidence set for its joint score; AVerImaTeC explicitly combines adequate evidence with a correct label. Neither rule licenses treating any relevant quotation as sufficient simply because a model predicts the expected answer. (fever.ai)

A paired temporal evidence fixture

The following is a **Hypothetical Example**. Names, dates, versions, and outcomes are invented. The fixed claim is: **“Mira Sen was chair of Alder Council throughout April 2024.”** The target interval starts on 2024-04-01 inclusive and ends on 2024-05-01 exclusive. The as-known cutoff is the end of 2024-05-04 in Coordinated Universal Time.

Table 2 records four different evidence dates. Revision identifies publication of changed content. Interval ends are excluded. “Unknown” labels eligibility; “unresolved” labels a verdict that the admitted evidence cannot establish.

Candidate	Asserted event interval	Publication or revision	Observation	Database recording	Version	Textual relation	Eligibility and verdict in as-known mode
A, contemporaneous	2024-04-01 to 2024-05-01	2024-05-01	2024-05-02	2024-05-03	A1	Supports throughout	Eligible; support if accepted alone.
B, stale paired passage	2024-01-01 to 2024-02-01	2024-05-01	2024-05-02	2024-05-03	B1	Same chair relation, wrong interval	Ineligible for April; unresolved if alone.
A, corrected source	2024-04-15 to 2024-05-01	2024-05-05	2024-05-06	2024-05-07	A2, corrects A1	Refutes throughout	Ineligible at cutoff; potentially usable retrospectively.
C, recurring occurrence	2023-04-01 to 2023-05-01	2024-05-01	2024-05-02	2024-05-03	C1	Prior annual occurrence	Ineligible for April 2024; unresolved if alone.
D, contradictory interval	2024-04-15 to 2024-05-01; not chair earlier	2024-04-20	2024-04-21	2024-04-22	D1	Refutes throughout	Eligible; unresolved with A1 if conflict lacks resolution.
E, future publication	2024-04-01 to 2024-05-01	2024-05-08	2024-05-09	2024-05-10	E1	Supports throughout	Ineligible at cutoff; potentially usable retrospectively.
F, missing anchor	“During April,” year unstated	Unverified	2024-04-25	2024-04-26	F1	Indeterminate	Unknown eligibility; unresolved if alone.

The A1/B1 pair holds the answer and publication freshness constant while changing the asserted interval. Returning “support” for both produces **correct-looking answer, failed evidence** for B1.

The A1/A2 pair tests **version leakage**. At the cutoff, A1 supports the claim under the source policy. A2 cannot retroactively become evidence available by the cutoff. A retrospective inquiry can use the correction and reach a different conclusion. Neither result proves A1 accurate; each states what preserved evidence supports under its mode. The version relation makes this distinction inspectable, consistent with provenance models that distinguish revision from the original entity. (www.w3.org)

The recurring pair tests occurrence identity despite matching entity and role. Wikidata documents a separate property for yearly recurring occurrences, illustrating why a recurring pattern and an individual occurrence need different temporal descriptions. (www.wikidata.org)

The remaining runs should be executed independently:

- **Contradiction run:** Admit A1 and D1 together and require the declared conflict policy.
- **Future-source run:** Replace A1 with E1; compare as-known and retrospective modes.
- **Missing-anchor run:** Admit F1 alone and require an explicit unknown reason.
- **Cutoff run:** Move the cutoff while preserving every captured version and claim.
- **Boundary run:** Replace “throughout” with “at some point” and treat that as a separately identified claim.

These variations make mistakes attributable. An unresolved verdict from missing admissible evidence is an abstention, not a false claim label. A contradiction can remain unresolved even when each candidate is temporally eligible.

Evaluation design

The proposed evaluation uses separate labeled sets for retrieval, eligibility, and verdicts. Evidence annotations should be completed independently of a model's answer wherever feasible. Otherwise an expected verdict can influence which timestamps annotators accept, undermining the intended test.

Retrieval coverage needs a defined corpus and reference set. The National Institute of Standards and Technology (NIST) retrieval definitions distinguish precision, the relevant fraction of retrieved documents, from recall, the retrieved fraction of relevant documents. For this audit, the reference set must identify temporally eligible evidence rather than merely topically related material. Open-web recall cannot be inferred from a small retrieved set without a defensible reference collection. (trec.nist.gov)

F1 is the harmonic mean of precision and recall. Micro averaging pools class counts, whereas macro averaging weights class scores equally. (scikit-learn.org) **Eligibility evaluation** uses candidate-level labels: eligible, ineligible, and unknown under a specified audit policy. Report the complete confusion matrix and per-class results. Precision and recall for the eligible class follow standard classification definitions, while macro averaging gives equal weight to labels rather than pooling their counts. Unknown should remain visible instead of being discarded as a convenient annotation exception. (scikit-learn.org)

Verdict evaluation uses claim-level support, refute, and insufficient labels derived from the admissible evidence set. An insufficient label describes what the set establishes, not a third state of worldly truth. Report results both with supplied reference evidence and with retrieved evidence. The former isolates reasoning more effectively; the latter exposes retrieval and eligibility failures as part of the end-to-end task. FEVER's evidence-conditioned score shows why label-only accuracy cannot fully describe this distinction. (fever.ai)

Abstention evaluation asks whether the system withholds answers where the admitted evidence is inadequate. Selective-classification research explicitly studies rejecting some predictions in exchange for higher accuracy, and SelectiveNet jointly learns prediction and rejection. Those findings motivate risk-coverage reporting; they do not validate a temporal abstention mechanism. (jmlr.org) (proceedings.mlr.press)

The evaluation report should include:

- **Corpus boundary:** Snapshot, permitted sources, and claim sampling procedure.
- **Annotation basis:** Raw timestamps, normalized intervals, and uncertainty reasons.
- **Retrieval result:** Eligible-reference recall and the size of the reference collection.
- **Eligibility result:** Confusion counts, class precision and recall, and unknown rate.
- **Verdict result:** Per-class metrics and aggregate results on a named claim set.
- **Selective result:** Answer coverage, accuracy on answered claims, and abstention errors.
- **Calibration result:** Observed correctness by declared confidence band, with sample counts.
- **Uncertainty:** Resampling unit, interval method, and any dependence among repeated claims.

Coverage curves must use the same fixed evaluation population as thresholds change. For confidence calibration, a high confidence score attached to an ineligible citation should remain an evidential failure even when the answer matches a label. NIST's framework calls for uncertainty measures and documented benchmark comparisons, providing a general reporting basis without prescribing this particular audit. (airc.nist.gov)

Data Analysis and Evidence

Published gains require explicit denominators

Allein's Table 2 supports direct arithmetic: (0.6417 - 0.5520) × 100 = **8.97 percentage points** for Micro F1; (0.4743 - 0.3239) × 100 = **15.04 points** for Macro F1. Dividing those differences by the baselines instead gives approximately **16.25%** and **46.43% relative gains**. The same paper reports RoBERTa changing from **0.6952/0.5532 to 0.6947/0.5739** in Micro/Macro F1, so the combined temporal configuration does not improve every metric or encoder. (aclanthology.org)

TemporalFC's abstract states **0.13 to 0.15 AUROC** improvement. (ris.uni-paderborn.de) However, its time-corrupted test comparison in Table 4 shows **0.65/0.65** against FactCheck's **0.52/0.53** on DBpedia124k/Yago3k, differences of **0.13/0.12**. This specific table comparison does not reproduce the abstract's full range. Table 5 separately gives MRR **0.70/0.83** against T-DYHE's **0.43/0.58**, yielding **0.27/0.25**. These values measure different tasks; averaging them would produce no meaningful common performance measure. (papers.dice-research.org)

ChronoFact's reported overall Macro F1 is **85.49 ± 0.53**, compared with **63.42 ± 0.60** for TACV. Its recurring-event subset scores are **80.89 ± 0.55** and **65.84 ± 5.04**, respectively. These are paper-reported comparisons on ChronoClaims, not a reproduced test of source publication, observation, or recording eligibility. Its missing insufficient-evidence class further limits any inference about abstention. (ijcai.org)

A reproducible hypothetical scoring calculation

The remainder of this section is a **Hypothetical Example**, using invented annotations. Suppose a closed reference corpus contains **20 eligible evidence items** for the evaluated queries, of which retrieval returns **12**. Retrieval recall is **12/20 = 60%**, following the relevant-items denominator rather than the number of retrieved passages. (trec.nist.gov)

Table 3 gives eligibility predictions for **30 retrieved candidates**, distinct from the claim-level verdict set below.

Gold eligibility	Predicted eligible	Predicted ineligible	Predicted unknown	Row total
Eligible	9	1	2	12
Ineligible	2	7	1	10
Unknown	1	2	5	8
Total	12	10	8	30

The eligible-class precision is **9/12 = 75%**, and its recall among retrieved candidates is **9/12 = 75%**. Accuracy across all eligibility classes is **21/30 = 70%**. Accepted eligible evidence reaches only **9/20 = 45%** of the full eligible reference set. The retrieval and eligibility scores therefore describe separate losses; either isolated 75% would hide the upstream shortfall. Precision and recall use their standard class-count definitions. (scikit-learn.org)

For an independent invented set of **20 claims**, let gold support/refute/insufficient counts be **8/6/6**. Predictions by gold row are support **6/1/1**, refute **1/4/1**, and insufficient **2/1/3**, with columns in that same order. Verdict accuracy is **13/20 = 65%**, and Macro F1 is approximately **0.639**, computed as the unweighted mean of class F1 values **12/17**, **8/12**, and **6/11**. (scikit-learn.org)

Treating predicted insufficient as abstention gives **15/20 = 75% answer coverage** and **10/15 = 66.7% accuracy on answered claims**. Of the **five abstentions**, **three** match gold insufficient, giving **60% abstention precision**; the system abstains on **3/6 = 50%** of gold-insufficient cases. These calculations are diagnostic examples, not evidence that any named method achieves these rates. A calibration claim would additionally require confidence predictions, which this invented fixture does not supply. The selective-classification literature motivates reporting the coverage-accuracy tradeoff rather than hiding abstention inside one aggregate. (jmlr.org)

Implications and Future Directions

The principal unresolved issue is **independent evidence-validity annotation**. Existing papers evaluate useful temporal reasoning tasks, but their headline metrics do not establish that source publication, version, observation, recording, and event validity were jointly labeled. ChronoClaims' absence of insufficient examples makes that limitation especially concrete; TSVer's annotated time frames address a different numerical evidence setting. (ijcai.org) (aclanthology.org)

Timestamp reliability needs its own uncertainty model. Crossref explicitly distinguishes record creation and later indexing modifications, and warns that publication dates can change or differ from record creation. DataCite separately defines creation of a digital object identifier (DOI) record and the resource's published field. A metadata update is therefore not evidence that an underlying event changed, nor proof that a newly displayed passage existed at the original publication date. (www.crossref.org) (support.datacite.org)

Historical source verification also remains bounded by capture coverage. RFC 7089 warns about sparse archived states. An archive gap should yield a precise statement about unavailable versions, rather than an assertion that no source ever existed. Similarly, a capture after the cutoff may preserve useful historical content without independently proving the same content was publicly available before it. (www.rfc-editor.org)

A practical research agenda should prioritize:

- **Clock annotation:** Independently label event time, publication, revision, observation, and recording.
- **Version pairs:** Preserve original and corrected text without overwriting the historical fixture.
- **Recurring events:** Distinguish occurrences that share entities, relation names, and seasonal language.
- **Partial intervals:** Retain granularity, uncertain endpoints, and alternative normalizations.
- **Claim stability:** Give changed temporal operators new identifiers instead of silently rewriting claims.
- **Domain transfer:** Repeat the audit across evidence modalities and source conventions.
- **Abstention tests:** Include genuinely insufficient and conflicting admitted evidence sets.
- **Reproducible reporting:** Publish fixtures, policies, confusion counts, and uncertainty calculations.

Cross-domain replication is a research requirement, not a result established by these examples. NIST's framework calls for documenting limits on generalizability beyond development conditions. The appropriate output is therefore a bounded evidence statement: which claims, which corpus, which clocks, which policy, and which unresolved cases. (airc.nist.gov)

Conclusion

Temporal fact verification becomes auditable when the claim interval and the evidence clocks remain explicit from retrieval through verdict. A publication date cannot substitute for the interval described by the text. An observation date cannot establish an earlier version's contents. A database history cannot determine whether its preserved assertion is true.

The proposed protocol asks evaluators to freeze the claim, declare historical or as-known mode, preserve the source version, and annotate textual relation separately from temporal eligibility. Provenance links make revision inspectable, but they do not remove the need to assess the assertion and the interval it concerns. (www.w3.org)

The paired fixture makes that distinction operational. A stale passage can produce an expected answer with inadmissible support. A corrected source can inform a later historical assessment without becoming evidence available at an earlier cutoff. A recurring event can match the entity and vocabulary while referring to the wrong occurrence. Conflicting eligible sources can warrant an unresolved result.

Evaluation should preserve those distinctions. Retrieval coverage identifies missing evidence; eligibility scoring tests admission decisions; verdict metrics measure conclusions on defined claim sets; and abstention analysis shows when a system withholds an answer. The established coverage-accuracy tradeoff supplies a reason to report selective performance explicitly. (jmlr.org)

A defensible result is consequently more than a correct label. It is a conclusion tied to a fixed proposition, interval, cutoff, preserved evidence, and declared policy, with uncertainty left visible when those materials do not establish the answer.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://aclanthology.org/2023.findings-eacl.13
2. https://ris.uni-paderborn.de/record/50479
3. https://ijcai.org/proceedings/2025/893
4. https://aclanthology.org/2023.findings-eacl.13.pdf
5. https://mariadb.com/docs/server/reference/sql-structure/temporal-tables/bitemporal-tables
6. https://www.w3.org/TR/prov-o
7. https://www.rfc-editor.org/rfc/rfc7089.html
8. https://jmlr.org/papers/v11/el-yaniv10a.html
9. https://fever.ai/task.html
10. https://fever.ai/2018/task.html
11. https://aclanthology.org/2025.emnlp-main.1519
12. https://docs.oracle.com/en/database/oracle/oracle-database/26/vldbg/time-based-info.html
13. https://learn.microsoft.com/en-us/sql/relational-databases/tables/temporal/overview?view=sql-server-ver17
14. https://www.crossref.org/documentation/retrieve-metadata/rest-api/tips-for-using-the-crossref-rest-api
15. https://www.mediawiki.org/wiki/API:Revisions
16. https://www.w3.org/TR/owl-time
17. https://www.postgresql.org/docs/current/rangetypes.html
18. https://datacite-metadata-schema.readthedocs.io/en/4.7/appendices/appendix-1/dateType
19. https://support.datacite.org/docs/what-is-the-difference-between-the-created-and-registered-date-in-the-datacite-rest-api
20. https://www.ibm.com/docs/en/ias?topic=tables-bitemporal
21. https://papers.dice-research.org/2023/ISWC_TemporalFC/public.pdf
22. https://aclanthology.org/2025.emnlp-main.1519.pdf
23. https://www.wikidata.org/wiki/Help:Dates
24. https://mementoweb.org/guide/quick-intro
25. https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml
26. https://trec.nist.gov/presentations/TREC9/overview/tsld014.htm
27. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html
28. https://proceedings.mlr.press/v97/geifman19a.html
29. https://airc.nist.gov/airmf-resources/airmf/5-sec-core

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.